

Structure of 311 Service Requests as a Signature of Urban Location

Lingjing Wang^{1,2}, Cheng Qian^{1,2}, Constantine Kontokosta^{1,2},
Stanislav Sobolevsky^{1,*}

¹ Center for Urban Science and Progress, New York
University, Brooklyn, New York, United States of America

² Tandon School of Engineering, New York University,
Brooklyn, New York, United States of America

* Correspondence should be addressed: sobolevsky@nyu.edu

Abstract

While urban systems demonstrate high spatial heterogeneity, many urban planning, economic and political decisions heavily rely on a deep understanding of local neighborhood contexts. We show that the structure of 311 Service Requests enables one possible way of building a unique signature of the local urban context, thus being able to serve as a low-cost decision support tool for urban stakeholders. Considering examples of New York City, Boston and Chicago, we demonstrate how 311 Service Requests recorded and categorized by type in each neighborhood can be utilized to generate a meaningful classification of locations across the city, based on distinctive socioeconomic profiles. Moreover, the 311-based classification of urban neighborhoods can present sufficient information to model various socioeconomic features. Finally, we show that these characteristics are capable of predicting future trends in comparative local real estate prices. We demonstrate 311 Service Requests data can be used to monitor and predict socioeconomic performance of urban neighborhoods, allowing urban stakeholders to quantify the impacts of their interventions.

Introduction

Cities can be seen as a complex system composed of multiple layers of activity and interactions across various urban domains; therefore, discovering a parsimonious description of urban function is quite difficult [1–4]. However, urban planners, policy makers and other types of urban stakeholders, including businesses and investors, could benefit from an intuitive proxy of neighborhood conditions across the city and over time [5–7]. At the same time, such simple indicators could provide not only valuable information to support urban decision-making, but also to accelerate the scalability of successful approaches and practices across different neighborhood and cities, as urban scaling patterns have become an increasing topic of interest [8–12]. As the volume and

heterogeneity of urban data have increased, machine learning has become a viable tool for enhancing our knowledge of urban space and in developing predictive analytics to inform city management and policy [1, 3, 13, 14].

The non-trivial challenge is to identify a consistent, quantifiable metric that provides comprehensive insights across multiple layers of urban operations and planning [17] and to locate readily-available data to support its implementation across a range of cities. Fortunately, urban data collected by various agencies and companies provide an opportunity to respond to this challenge [18, 19]. In the age of ubiquitous digital media, numerous aspects of human activity are being analyzed by means of their digital footprints, such as mobile call records [20–26], vehicle GPS traces [27], bank card transactions [28–30], payment patterns [31–33], smart card usage [34–37], or social media activity [38–42]. Such data sets have been successfully applied to investigate urban [43] and regional structure [23, 44], land use [45, 46], financial activities [47], mobility [48, 49], or well-being [35, 50].

However, one of the major limitations to widespread adoption of such analytics in the practice of urban management and planning is the extreme heterogeneity of the data coverage: different types of data are available for different areas and periods of time, which undermine efforts to develop universal and reliable analytic approaches. Privacy considerations are another significant issue that create additional practical and legal obstacles, restricting data access and preventing their use out of a concern for confidentiality [51–54].

Increasingly, cities are introducing systems to collect service requests and complaints from their citizens. These data, commonly referred to as 311 requests, reflect a wide range of concerns raised by city residents and visitors, offering a unique indicator of local urban function, condition, and service level. In many cities, 311 requests are publicly available through city-managed open data platforms as part of a broader movement in local government to increase transparency and good governance [55]. Although potentially biased by the self-reported nature of the requests and complaints, these data provide a comparable measure of perceived local quality of life across space and time.

In this article, we develop a method for classifying urban locations based on the categorical and temporal structure of 311 Service Requests for a given neighborhood, exploring whether these spatio-temporal patterns can reveal characteristic signatures of the area. For New York, Boston, and Chicago, we present applications of this new urban classifier for predicting socioeconomic and demographic characteristics of a neighborhood and estimating the economic performance and well-being of a defined spatial agglomeration. The paper begins with a discussion of the data and methodology, followed by specific use cases relating to demographics and real estate values, and concluding with opportunities for future research.

1 Materials and Resources

1.1 The 311 data

311 service request and complaint data are being collected across more than 30 cities in the United States, including New York, Boston and Chicago. Through the 311 system, local government agencies offer non-emergency services to residents, visitors and businesses and respond to reported service disruptions, unsafe conditions, or quality-of-life disturbances. These 311 service requests and complaints cover a wide range of concerns, including, but not limited to, noise, building heat outages, rodent sightings, etc. Thus, these data serves as an extremely useful resource in understanding the delivery of critical city services and neighborhood conditions.

We explore the 311 datasets from New York, Chicago, and Boston as major urban

centers where 311 systems are in place and commonly used. We consider a time frame between 2012 and 2015 during which the data are available for all three cities selected. In table 1, we provide descriptive statistics of the data. Note that the number of total requests has been increasing from 2012 to 2015 in each city. Conceivably, the number of requests in New York City (which now approaches 2 million per year) is higher than the others because of its population size. However, Boston has a substantially smaller number of requests compared to the similar-sized city of Chicago, which shows the discrepancies in the use of the system across cities. Unfortunately, each city uses a different complaint/request coding convention, thus there is little consistency in the classification of particular complaint types. This fact raises certain difficulties for analysis between cities, a common challenge in comparative urban analytics given the lack of data standardization. For example, in 2015, New York City’s 311 data are categorized into 182 types, where Chicago has only 12. Even within a particular city, request categories are subject to change over time, especially in NYC where only approximately 70% of the entire service request activity belong to common categories present in all four years. Additional adjustments are needed to re-classify complaint types into standardized categories across the different cities and over the time period of the analysis.

The original data set provided by 311 Services contains one record for each customer’s call. For most cities, these records include information such as: service request type, service request open/close time and date and location(longitude and latitude). Therefore, for any given time period and area(census tract area/zipcode area), we can aggregate the 311 service requests and group by type.

Year	New York City		
	Total Requests	Requests Categories	Share of common categories’ activity
2012	1414392	165	0.69
2013	1431729	162	0.69
2014	1654913	179	0.73
2015	1806560	182	0.73
Year	Chicago		
	Total Requests	Requests Types	Share of common categories’ activity
2012	478532	13	0.85
2013	507956	14	0.82
2014	515258	14	0.82
2015	568576	12	0.9
Year	Boston		
	Total Requests	Requests Types	Share of common categories’ activity
2012	92855	155	1
2013	112727	165	0.99
2014	112785	183	0.96
2015	161498	180	0.83

Table 1. General properties of the 311 data for NYC, Chicago and Boston

1.2 Demographic and socio-economic data

As we are attempting to use 311 data as a proxy for the socioeconomic characteristics and real estate values of urban neighborhoods, ground-truth data are needed to train and validate our models. For socioeconomic and demographic features, we use data from the U.S. Census 2014 American Community Survey (ACS). For real estate values, we collect housing price data from the online real estate listing site Zillow. Both are

described below.

1.2.1 2014 census data

The 2014 ACS contains survey data on a number of socioeconomic and demographic variables, at the spatial aggregation of the Census Block. For this analysis, we have selected common features representing important phenomena in population diversity, education, and income and employment. For example, our selection covers the number of population in the following categories: "Non-Hispanic White", "African-American", "Asian", "High school degree", "College degree", "Graduate degree", "Uninsured ratio", "Unemployment ratio", "Poverty ratio", and mean for "Income (all)", "Income of No Family", "Income of Families" and "Income of Households".

One important consideration is the level of spatial aggregation for this analysis. Having considered zip code, census tract and census block, we decided to proceed with census tracts providing the best trade-off between spatial granularity, in terms of having a sufficient number of sub-areas within each city, and having a statistically significant sample of 311 complaints for each areal unit. In Boston and Chicago, there are too few zipcodes within in each city to create a useful sample, and there is not a significant density of 311 complaints at the census block level (please refer to SI4 for details). In addition, given the survey methodology of the ACS data, census block data include non-trivial margins-of-error for each variable.

1.2.2 Zillow Housing price

One important indicator of local economic conditions is housing prices [56]. We utilize Zillow housing price data that contain monthly average residential real estate sales prices by zip code. Although housing prices are a lagging indicator of neighborhood economic strength, since recorded sales occur as much as two to more than six months after a contract is signed, we use these values as one of the targets for our 311 predictions. Our spatial level of analysis will be the zipcode, rather than census tracts, given the coverage area of the Zillow aggregate data.

1.2.3 Normalization method and some notations

In order to better compare various areas, the census data need to be normalized. Take income per capita and population with bachelor degree for example. Firstly, these two features have different measurement units (dollars versus number of people). Secondly, this number can be affected by the area's total population. For an area with high population, there should be a higher possibility to have higher population with bachelor degree. Therefore, the normalization process is important in order to compare different features and different areas with heterogeneous population. For our analysis, we normalize our census tract data set in the following way.

Let p_i be the total population in census tract, while v_i denotes one feature recorded in the same census tract i , for example, "the total population who holds graduate degree in census tract i ". Next, we normalize it by defining

$$y_i = \frac{v_i p_i}{\sum_{j \in \Omega} v_j p_j}$$

We define Ω as a set of all census tracts in New York City.

In section 2, we use 311 complaint frequency categorized by census tracts to cluster and investigate the difference in local socioeconomic features y . In section 3 we use machine learning regression models to predict these features y using normalized 311 data .

2 Classification based on 311 service categories

In order to get initial insights on the usage of 311 service across the considered cities, we define for each census tract a 311 service request signature - a vector of the relative frequencies of 311 requests of different types. Specifically, let the total number of service requests of each type t within an area a be $s(a, t)$ and let $s(a) = \sum_t s(a, t)$ be the total number of service requests in the area a . Then a vector $S(a) = (s(a, t)/s(a), t = 1..T)$, where T is the total number of service request types, serves as a signature of the location’s aggregated 311 service request behavior. The vector S highlights the primary reasons for service requests or complaints in the specific area, as well as allowing for straightforward comparison across tracts and cities.

Signatures $S(a)$ serve as unique characteristics of each location a , and we would expect similar spatio-temporal patterns to emerge in 311 service requests across a city or cities. Our hypothesis here is that these similarities also suggest similarities in the socioeconomic characteristics of the areas. In order to explore this further, we apply a k -means clustering approach to the set of multi-dimensional vectors $S(a)$. In order to ensure we get an optimal clustering we run the k -mean 100 times, selecting the best solution in terms of cumulative square sum of distances from centroids.

One crucial step in this approach is to pick up an appropriate number of clusters to consider. For that purpose we have evaluated the clustering model with both Silhouette method and Elbow method. While different methods give a slightly different optimal number of clusters for the cities in our sample, in most cases it is within a range of two to four clusters. Given the socioeconomic diversity across neighborhoods in the selected cities, we determine that a minimum of four clusters is an appropriate value. Readers can find more details in SI.

We consider NYC first. In Figure 1, we see below with approximate 2000 census tracts divided into four clusters based on our clustering results. Midtown Manhattan, downtown Brooklyn and several outliers such as JFK and LGA airports belong to cluster 1; Staten Island and eastern Brooklyn/Queens constitute cluster 2; Northern Manhattan, the Bronx, and central Brooklyn are included in cluster 3; and Southern Brooklyn, Flushing and some eastern parts of Bronx comprise cluster 4.

In order to evaluate how different each cluster is with respect to the nature of 311 service requests, (see figure 2) we present the distribution of top service requests over the four clusters. We observe clear variation in this distribution. For example, complaints/requests within cluster 1 more often report noise concerns than others, cluster 2 experiences more issues relating to residential heating, cluster 3 has the highest relative complaints about blocked driveways, while cluster 4 reports concerns about street conditions.

Similarly, we repeat the same clustering process for Chicago and Boston and the clustering results for census tracts in those cities are shown in Figure 3.

3 Socioeconomic features among clusters

Given knowledge of the local spatial contexts for the analyzed cities, the clusters that emerge make certain intuitive sense. However, in order to quantitatively address the hypothesis formulated in the previous section - that similarities in local 311 service request signatures also imply similarities in the socioeconomic profiles of those areas - here we summarize and analyze the socioeconomic characteristics for each of the discovered clusters.

Recall that thus far the clustering results are obtained based on the 311 service requests frequency alone with no socioeconomic information considered. Next we summarize 14 socioeconomic features and compare the normalized mean level for each

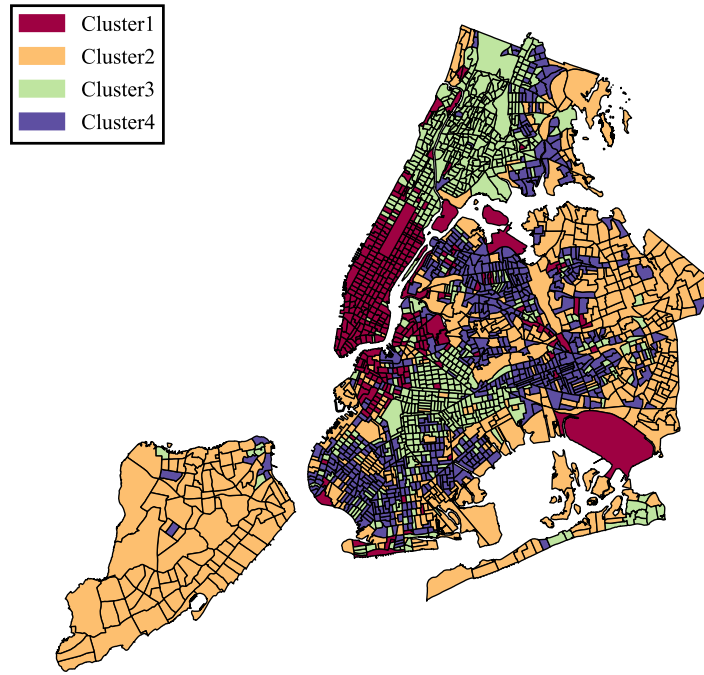


Figure 1. Classification of urban locations based on the categorical structure of the 311 requests.

feature in each of the considered clusters. The results for our three cities are presented in the radar plots in Figures 4-6. From the output, we can see that the socioeconomic features among the defined clusters are quite distinctive.

Take NYC for example:

- Education and Income: People with higher levels of education (with graduate degree and above) are found in cluster 1, which, as expected, also has highest income level. Cluster 3 appears to show the opposite results.
- Racial diversity: There are above average concentrations of Non-Hispanic Whites living in clusters 1 and 2, of Asian origin in cluster 4, and African-American populations in cluster 3.

Similarly we have (for both Chicago and Boston):

- Cluster 1 has the highest income and education level, while cluster 3 is the lowest.
- Cluster 2 is predominantly Asian and African-Americans, while Non-Hispanic Whites tend to live in clusters 1 and 4.

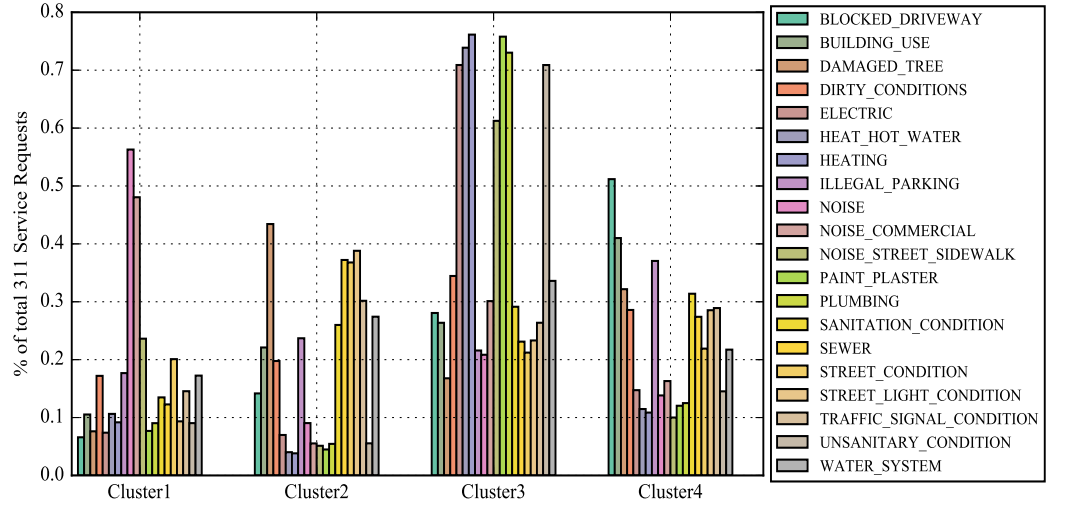
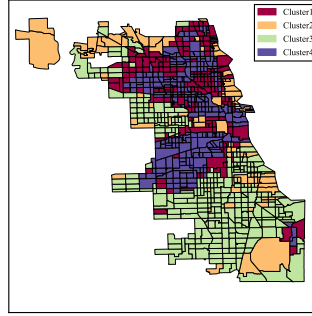
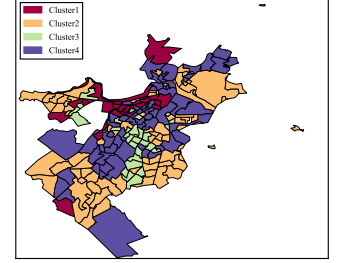


Figure 2. Top 20 requests distribution among clusters.



(a) Chicago



(b) Boston

Figure 3. Classification of urban locations based on the categorical structure of the 311 service requests for Chicago and Boston.

The observations above provide some evidence for our hypothesis, revealing links between socioeconomic features and 311 service request data structure. Indeed, while the clustering is performed based on the 311 data alone, the socioeconomic features happen to be quite distinctive among the clusters. Of course this only reveals the existence of a certain relation in principle, which might not be that practical. However this gives rise to another hypothesis - can one use 311 service request data to actually model socioeconomic features at the local scale?

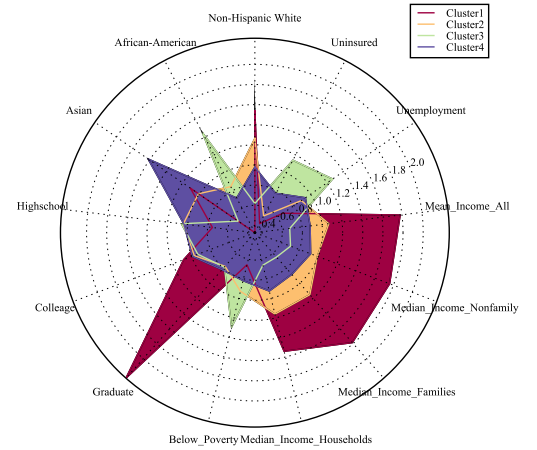
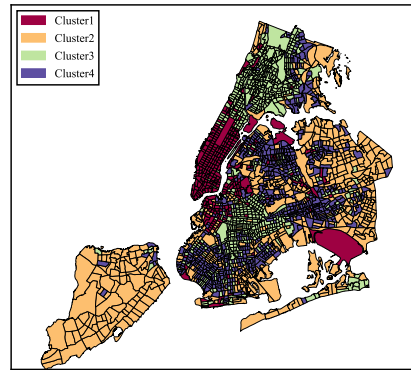


Figure 4. Normalized ratio of socioeconomic features among clusters in New York

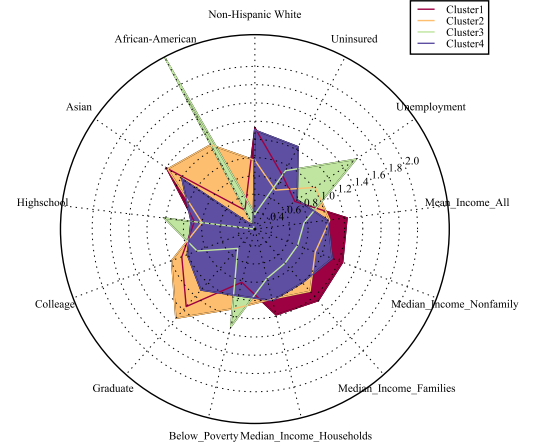
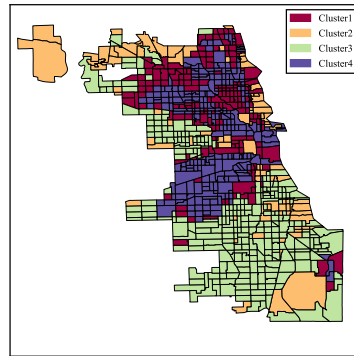


Figure 5. Normalized ratio of socioeconomic features among clusters in Chicago

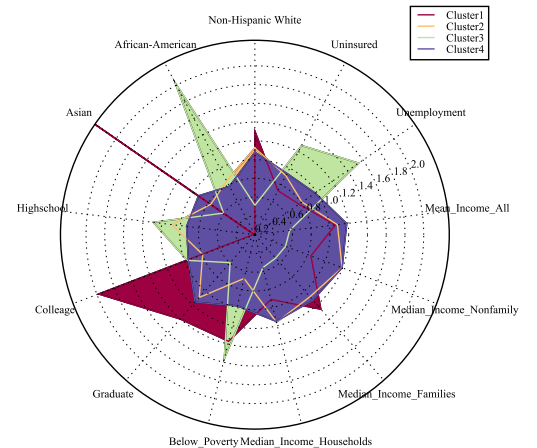
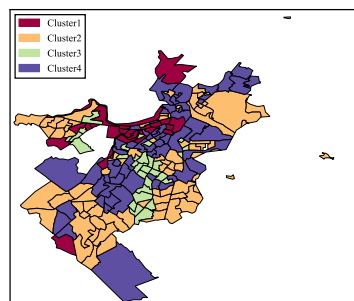


Figure 6. Normalized ratio of socioeconomic features among clusters in Boston

4 Modeling the socioeconomic features

We find that 311 service request signatures allow the city to be divided into clusters based on distinctive patterns of socioeconomic characteristics. Following this, we explore whether 311 service requests can be used to model these socioeconomic patterns. Such a model could be useful as socioeconomic data are often unavailable or inconsistent at a given spatio-temporal scale, and therefore having a proxy derived from a model based on regularly-updated open data could have considerable potential for city operations and neighborhood planning.

We train regression models over the relative frequencies of 311 service requests of each type in each census tract in order to estimate the selected socioeconomic features described in subsection 1.2.1. The service requests frequencies $s(a, t)/s(a)$ (components of the signature vectors) constitute our feature space, including 179 different features in the case of NYC, across 2000 census tracts following the data cleaning/filtering process.

We consider six target variables including income per capita, percentage of residents with a graduate degree, percentage of unemployed residents, percentage of residents living below the poverty level, as well as demographic characteristics including percentage of Non-Hispanic White and African-American populations.

The objective of the modeling is to use partial information about the target variables defined in a certain part of the city to train the model so that it can explain the target variables over the rest of the city.

For the purpose of a comprehensive model evaluation we use a cross-validation procedure. We try several models including Lasso [57], Neural Networks with regularisation (NN) [58–60], Random Forests Regression (RF) [61] and Extra Trees Regression (ETR) [62, 63].

For each model, we treat the different set of hyper parameters as different models. For Neural Networks, we try 5, 10, 20, 40 hidden unites and for each hidden unit, we try penalization lambda for 0.0005, 0.005, 0.05, 0.5. As to the learning process, we use mini batch size 20 and we use the following learning rate and epochs: (0.1,100), (0.05,200), (0.01,500), (0.005,1000). For RF and ETR, we use 500 trees(since increasing trees does not help) and try maximum leaf nodes: 10, 20, 30, ... 100. In total we have 64 sets of hyper parameters for NN and 10 for RF and 10 for ERF.

More details on the model selection process is presented in SI 3. Generally speaking, we select the final model with suitable hyper parameters with the help of cross-validation. We divide the data set into training and testing set by the ratio 7:3. As described above, we have 84 different models. For each model, we randomly divide our training set into training and validation sets and train the model on the training set and report the R-squared on the validation set. We repeat this process 20 times for each model and get the average R-squared. We pick the best model and use it for prediction on the test set. Finally, we report the out-of-sample R-squared in Table 2. Generally speaking, RF and ETR usually give us best performance based on R-squared.

The resulting out-of-sample R-squared for the models selected are summarized in Table 2. We consider this modeling result important because:

- it indicates that a relationship exists between 311 request signature and the local socioeconomic features of each area;
- it enables possible prediction and estimation of other local socioeconomic features by using 311 requests data, particularly those features for which data are collected at low temporal frequency, such as Census data; and,
- it can be easily scaled by geographic aggregation for various research, operational, or planning purposes.

City	White/European	Afro-American	Graduate Degree	Income per cap	Below Poverty	Unemployment
NYC	0.54	0.50	0.48	0.70	0.44	0.26
Chicago	0.76	0.85	0.45	0.55	0.52	0.65
Boston	0.54	0.68	0.26	0.62	0.63	0.36

Table 2. Out of Sample R-squared

5 Prediction of the real estate prices

Following our previous analysis, we attempt to understand the practical applicability of the prediction models. Although the findings above once again highlight a strong relation between 311 service request data and socioeconomic context of urban locations, this by itself has limited practical implications except for filling gaps in the data availability. In this section we show that 311 service request data could be also used to predict future socioeconomic variations, which may have more important practical implications for urban analytics.

As an example, consider the annual average sale price of housing per square foot in different neighborhoods of NYC as the target variable for our prediction. Our housing price is reported by Zillow at the zip code level; therefore, we rescale our 311 service request frequencies to this spatial aggregation.

To match available housing price data, we only include those 311 service categories that were recorded consistently between 2012 and 2015. New York City has 145 of such categories, covering about 70 percent of total service requests.

The target variable is updated annually and is available for each year from 2012 to 2015. The Zillow data cover 112 of the 145 zip codes in New York City where the density and frequency of 311 requests is sufficient to satisfy the filtering procedure described in the Data section. Thus, our sample for this prediction is based on data from 112 zipcodes.

We do not attempt to predict the absolute level of prices, but changes over time relative to the NYC mean. Our output therefore indicates how much more (less) expensive the housing price in a given zip code area is going to be compared to the average relative increase (decrease) in housing prices across NYC from the previous year. This way we define a new log-scale target variable $Y^t(z)$ in year t as

$$Y^t(z) = \log(P^t(z)/P_{mean}^t)$$

where $P^t(z)$ is the average price per square foot in zip code z during the year t , while P_{mean}^t is the average price per square foot across the entire city during the year t , estimated as the mean of $P^t(z)$ for all the locations z weighted by residential population of the locations used as a proxy for the locations' size.

We begin by modeling the output variable Y^{2015} . We train the model using 2012 and 2013 data (both - features and output variable) over the entire NYC and use 2014 data for tuning hyper-parameters, then apply it to 2015 using the features defined based on 2015 service requests. To reiterate, the feature space as before consists of the relative service requests frequencies $s(a, t)/s(a)$, but now including only 145 categories of service requests, while the number of observations is 112 zip codes.

We subsequently train four different machine learning regression models as before: Lasso [57], Neural Networks with regularisation (NN) [58–60], Random Forests Regression (RF) [61] and Extra Trees Regression (ETR) [62, 63].

The results are reported in table 3 (we also include Boston and Chicago here just for comparison, although the number of zip codes in these cities is much smaller and thus the model becomes less significant).

As one can see from the table 3, we achieve reasonable predictive power, especially with RF and ETR approaching R^2 values of 0.80 for all three cities.

Models	NYC	Chicago	Boston
Lasso	0.49	0.57	0.38
NN(Regularized)	0.70	0.65	0.68
RF	0.78	0.81	0.79
ETR	0.79	0.90	0.83

Table 3. Out of Sample R-squared

However, note that modeling housing prices in 2015 is not our objective here, since a simplified model $Y^{2015} = Y^{2014}$ would achieve better results given the serial correlation in the time series and the relatively small year-to-year variation in price levels. Instead, we rather focus on the model's ability to predict the magnitude and direction of those fluctuations, forecasting price trends at the zip code level.

Let $Y_P^t(z)$ be the predicted value of $Y^t(z)$. We define $D(z) = Y^{2015}(z) - Y^{2014}(z)$ as the actual tendency of relative real estate prices in the zip code z and $D_P(z) = Y_P^{2015} - Y_P^{2014}$ as the predicted tendency of comparative housing price.

We classify the 112 zip codes of NYC into three groups based on the predicted tendency strength D_P^{2015} :

$G_{Positive} = \{z : D_P^i > m \cdot \sigma(D_P), \text{ where } i = 1, 2, \dots, 112\}$: group of areas with strong positive tendency;

$G_{Negative} = \{z : D_P^i < -m \cdot \sigma(D_P), \text{ where } i = 1, 2, \dots, 112\}$: group of areas with strong negative tendency;

$G_{Neutral} = \{z : -m \cdot \sigma(D_P) < D_P(z) < m \cdot \sigma(D_P), i = 1, 2, \dots, 112\}$: group of areas with close to neutral tendency,

where m is a certain threshold and $\sigma(D_P)$ indicates the standard deviation of $D_P(z)$.

Additionally we classify the zip codes based on the actual tendency strength, i.e. let us introduce $G'_{Positive}, G'_{Negative}, G'_{Neutral}$ in the same way as above but replacing the estimated $D_P(z)$ with the real $D(z)$ in the corresponding. In this way, compared to defining strong tendency using predicted results, we define strong tendency by the real values and then test the performance of our model by the following indicators.

For each group $G_{Positive}, G_{Negative}, G_{Neutral}$, we calculate its the normalized population weighted average value of actual $D(z)$ using the following formulae:

$$\begin{aligned}\bar{D}_{Positive} &= \left(\frac{\sum_{i \in G_{Positive}} D(z) \cdot N(z)}{\sum_{z=1}^{112} D(z) \cdot N(z)} \right) / \sigma(D(z)), \\ \bar{D}_{Negative} &= \left(\frac{\sum_{i \in G_{Negative}} D(z) \cdot N(z)}{\sum_{z=1}^{112} D(z) \cdot N(z)} \right) / \sigma(D(z)), \\ \bar{D}_{Neutral} &= \left(\frac{\sum_{i \in G_{Neutral}} D(z) \cdot N(z)}{\sum_{z=1}^{112} D(z) \cdot N(z)} \right) / \sigma(D(z)),\end{aligned}$$

where $N(z)$ is the population of the zip code z . Similarly for each of the groups $G'_{Positive}, G'_{Negative}, G'_{Neutral}$ we calculate the average prediction

$$\begin{aligned}\bar{D}'_{Positive} &= \left(\frac{\sum'_{i \in G'_{Positive}} D_P(z) \cdot N(z)}{\sum_{z=1}^{112} D_P(z) \cdot N(z)} \right) / \sigma(D(z)), \\ \bar{D}'_{Negative} &= \left(\frac{\sum'_{i \in G'_{Negative}} D_P(z) \cdot N(z)}{\sum_{z=1}^{112} D_P(z) \cdot N(z)} \right) / \sigma(D(z)), \\ \bar{D}'_{Neutral} &= \left(\frac{\sum'_{i \in G'_{Neutral}} D_P(z) \cdot N(z)}{\sum_{z=1}^{112} D_P(z) \cdot N(z)} \right) / \sigma(D(z)),\end{aligned}$$

The values of those quantities for different values of the threshold ($m = 0.15$, example of a very loose threshold classifying most of the predictions as strong, $m = 0.35, 0.65, 1$) are reported in the Tables 4 and 5 and we can see consistent inequalities

$$\overline{D}_{Positive} > 0 > \overline{D}_{Negative}$$

and

$$\overline{D}'_{Positive} > 0 > \overline{D}'_{Negative}$$

holding for all the values of the threshold m , which means that our predicted trend directions are consistent with the real trends on average.

Moreover, we compare the signs of the predicted values of $D_P(z)$ for the strong predicted trends $G_{Positive} \cup G_{Negative}$ vs the ground-truth $D(z)$, as well as the actual values $D(z)$ for the strong actual trends $G'_{Positive} \cup G'_{Negative}$, reporting the accuracy ratio of predicting the correct trend direction for strong actual trends and the accuracy ratio for having strong predicted trends to reveal correct trend directions ($D_P(z)D(z) > 0$). Those indicators are listed in Table 4 and Table 5 demonstrating the model's performance.

From Tables 4 and 5, we see that, for around 40 percent of strongest tendency observations or predictions ($m=0.65$), our prediction accuracy of a trend direction is higher than 80 percent compared to around 43/62(69%) percent random guess baseline model in Table 4 and 31/51(60.7%) baseline in Table 5. Moreover, in Table 4, we see that when the threshold m increases from 0.15 to 0.65, the accuracy ratio of prediction goes up from 70 percent to 82 percent, meaning that the stronger the actual trend, the more likely to achieve correct prediction. In Table 5, we see that while m increases from 0.15 to 1, the accuracy ratio of prediction goes up from 72 percent to 90 percent, hence the stronger the predicted trend, the more accurately our prediction reflects the reality.

Threshold	m=0.15			m=0.35		
+/-:Strong Positive/Negative	+	-	Neutral	+	-	Neutral
Number of Observations	23	75	14	20	62	30
$\overline{D}'_{Positive}/\overline{D}'_{Negative}/\overline{D}'_{Neutral}$	134.57	-84.28	-3.75	148.60	-95.41	-7.97
Accuracy for Strong P/N	0.7			0.72		
Threshold	m=0.65			m=1		
+/-:Strong Positive/Negative	+	-	Neutral	+	-	Neutral
Number of Observations	19	43	50	14	24	74
$\overline{D}'_{Positive}/\overline{D}'_{Negative}/\overline{D}'_{Neutral}$	156.73	-114.82	-24.5	179.69	-137.11	-32.56
Accuracy for Strong P/N	0.82			0.77		

Table 4. Accuracy of discovering actual strong relative real estate price trends by the predictive model

The results presented in this section demonstrate that the 311-based model can indeed predict future fluctuations of socio-economic characteristics, including real estate price trends. This serves as an initial proof of concept for multiple potential urban applications using 311 data as a proxy for local socio-economic conditions.

Conclusions

A quantitative understanding of urban neighborhoods can be quite challenging for urban planners and policy-makers given significant gaps in the spatial and temporal resolution of data and data collection modalities. However, this subject is crucial for urban planning and decision making, as well as for the study of urban economic and

Threshold	m=0.15			m=0.35		
+/-:Strong Positive/Negative	+	-	Neutral	+	-	Neutral
Number of Observations	43	58	11	32	42	38
$\overline{D}_{Positive}/\overline{D}_{Negative}/\overline{D}_{Neutral}$	22.61	-75.99	-4.56	42.23	-71.18	-40.78
Accuracy for Strong P/N	0.72			0.77		
Threshold	m=0.65			m=1		
+/-:Strong Positive/Negative	+	-	Neutral	+	-	Neutral
Number of Observations	20	31	61	15	12	85
$\overline{D}_{Positive}/\overline{D}_{Negative}/\overline{D}_{Neutral}$	44.93	-70.55	-29.83	110.80	-76.29	-41.17
Accuracy for Strong P/N	0.83			0.90		

Table 5. Accuracy of the correspondence of the predicted strong relative real estate price trends to the actual ones

neighborhood change. In this paper, we provide an approach to quantify local signatures of urban function via 311 service request data collected in various cities across the US. These datasets, which can be easily scaled by spatial (zip code, census tracts/blocks, etc.) and temporal level of aggregation, are open to the public and updated regularly. Importantly, we demonstrate consistent relationships between socioeconomic features of urban neighborhoods and their 311 service requests.

For all three cities analyzed - New York City, Boston and Chicago - we demonstrate how clustering of census tracts by the relative frequency vectors of different types of 311 requests reveal distinctive socioeconomic patterns across the city. Moreover, those frequency vectors allow us to train and cross-validate regression models successfully explaining selected socioeconomic features, such as education level, income, unemployment and racial composition of urban neighborhoods. For example, the accuracy of the model explaining local average income in NYC is characterized by a R-squared value of 0.7, while Extra Trees Regression results in a 0.9 out of sample R-squared in explaining housing prices in Chicago (although this must be considered with respect to the smaller sample size). Finally, we illustrate the predictive capacity of the approach by training and validating the model to detect comparative average real estate price trends for zip codes in New York City.

In the nascent field of urban science and more traditional disciplines of economics and urban planning, there is increasing attention on how data collected by cities can be combined with novel machine learning approaches to yield insight for researchers and policy-makers. It is possible that such data can be used to better understand the dynamics of local areas in cities, and support more informed decision-making. In addition, it is conceivable that a set of efficient instrumental variables based on widely-available 311 data can be used to replace survey-based socioeconomic statistics at spatio-temporal scale where such official survey data is non-existent or inconsistent, thus broadening opportunities for urban analytics.

Acknowledgments

The authors would like to thank Brendan Reilly and other colleagues at the Center For Urban Science And Progress at NYU for stimulating discussions which helped further shaping this research and manuscript.

References

1. Michael Batty. The size, scale, and shape of cities. Science, 319(5864):769–771, 2008.
2. Luís MA Bettencourt, José Lobo, Deborah Strumsky, and Geoffrey B West. Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities. PLoS ONE, 5(11):e13541, 2010.
3. Luís MA Bettencourt. The origins of scaling in cities. Science, 340(6139):1438–1441, 2013.
4. Elsa Arcaute, Erez Hatna, Peter Ferguson, Hyejin Youn, Anders Johansson, and Michael Batty. City boundaries and the universality of scaling laws. arXiv:1301.1674, 2013.
5. Constantine E Kontokosta. The quantified community and neighborhood labs: A framework for computational urban science and civic technology innovation, 2015.
6. Oded Maimon and Lior Rokach. Data mining and knowledge discovery handbook. 2010.
7. A. M. Townsend. Smart cities: Big data, civic hackers, and the quest for a new utopia. New York: WW Norton and Company, 2013.
8. Lisa M Powell, Sandy Slater, Donka Mirtcheva, Yanjun Bao, and Frank J Chaloupka. Food store availability and neighborhood characteristics in the United States. Preventive Medicine, 44(3):189–195, 2007.
9. Luís MA Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B West. Growth, innovation, scaling, and the pace of life in cities. Proceedings of the National Academy of Sciences, 104(17):7301–7306, 2007.
10. Sergio Albeverio, Denise Andrey, Paolo Giordano, and Alberto Vancheri. The dynamics of complex urban systems. Physica, Heidelberg, 2008.
11. Stanislav Sobolevsky, Izabela Sitko, Sebastian Grauwin, Rem Tachet Des Combes, Bartosz Hawelka, Juan Murillo Arias, and Carlo Ratti. Mining urban performance: Scale-independent classification of cities based on individual economic transactions. In Big Data Science and Computing, 2014 ASE International Conference on, May 27-31, Stanford University, page 10, 2014.
12. Stanislav Sobolevsky, Iva Bojic, Alexander Belyi, Izabela Sitko, Bartosz Hawelka, Juan Murillo Arias, and Carlo Ratti. Scaling of city attractiveness for foreign visitors through big data of human economical and social media activity. In 2015 IEEE International Congress on Big Data, pages 600–607. IEEE, 2015.
13. J. A. Nelder and R. J. Baker. Generalized linear models. Encyclopedia of Statistical Sciences, 2006.
14. B Bolstad, R Irizarry, M Astrand, and T Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19(2):185—193, 2003.
15. J. MacQueen. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1: Statistics:281–297, 1967.

-
16. Peter Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20:53–65, 1987.
 17. S. Allwinkle and P. Cruickshank. Creating smart-er cities: An overview. Journal of urban technology, 18:1–16, 2011.
 18. Michael Batty. Smart cities, big data. Environment and Planning-Part B, 39, 2012.
 19. Bettencourt Luís M.A. The uses of big data in cities. Big Data, 2:12–22, 2014.
 20. Fabien Girardin, Francesco Calabrese, Filippo Dal Fiore, Carlo Ratti, and Josep Blat. Digital footprinting: Uncovering tourists with user-generated content. Pervasive Computing, IEEE, 7:5276, 2008.
 21. M.C. González, C.A. Hidalgo, and A.-L. Barabási. Understanding individual human mobility patterns. Nature, 453:779–782, 2008.
 22. D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft. Recommending social events from mobile phone location data. In Data Mining (ICDM), 2010 IEEE 10th International Conference on, pages 971–976, 2010.
 23. Stanislav Sobolevsky, Michael Szell, Riccardo Campari, Thomas Couronné, Zbigniew Smoreda, and Carlo Ratti. Delineating geographical regions with networks of human interactions in an extensive set of countries. PloS ONE, 8(12):e81707, 2013.
 24. Alexander Amini, Kevin Kung, Chaogui Kang, Stanislav Sobolevsky, and Carlo Ratti. The impact of social segregation on human mobility in developing and industrialized regions. EPJ Data Science, 3(1):6, 2014.
 25. Jonathan Reades, Francesco Calabrese, Andres Sevtsuk, and Carlo Ratti. Cellular census: Explorations in urban data collection. Pervasive Computing, IEEE, 6:30–38, 2007.
 26. Constantine E Kontokosta and Johnson N. Urban phenology: Toward a real-time census of the city using wi-fi data. Computers, Environment, and Urban Systems, 2016.
 27. Paolo Santi, Giovanni Resta, Michael Szell, Stanislav Sobolevsky, Steven Strogatz, and Carlo Ratti. Quantifying the benefits of vehicle pooling with shareability networks. Proceedings of the National Academy of Sciences, 111(37):13290–13294, 2014.
 28. S Shen, A Sam, and E Jones. Credit card indebtedness and psychological well-being over time: Empirical evidence from a household survey. Journal of Consumer Affairs, 48(3):431–456, 2014.
 29. B Scholnick, N Massoud, and A Saunders. The impact of wealth on financial mistakes: Evidence from credit card non-payment. Journal of Financial Stability, 9(1):26–37, 2013.
 30. Stanislav Sobolevsky, Izabela Sitko, Remi Tachet des Combes, Bartosz Hawelka, Juan Murillo Arias, and Carlo Ratti. Cities through the prism of people’s spending behavior. PloS one, 11(2):e0146291, 2016.

-
31. Willem C. Boeschoten. Cash management, payment patterns and the demand for money. The Economist, 146(1):117–142, 1998.
 32. David Bounie and Abel Francois. Cash, check or bank card? The effects of transaction characteristics on the use of payment instruments. SSRN Scholarly Paper, (ID 89179), 2006.
 33. Celia Ray Hayhoe, Lauren J. Leach, Pamela R. Turner, Marilyn J. Bruin, and Frances C. Lawrence. Differences in spending habits and credit use of college students. Journal of Consumer Affairs, 34(1):113–133, 2008.
 34. M. Bagchi and P.R. White. The potential of public transport smart card data. Transport Policy, 12(5):464–474, 2005.
 35. Neal Lathia, Daniele Quercia, and Jon Crowcroft. The hidden image of the city: Sensing community well-being from urban mobility. In Judy Kay, Paul Lukowicz, Hideyuki Tokuda, Patrick Olivier, and Antonio Krüger, editors, Pervasive Computing, volume 7319 of Lecture Notes in Computer Science, pages 91–98. 2012.
 36. Philip K. Chan, Wei Fan, Andreas L. Prodromidis, and Salvatore J. Stolfo. Distributed data mining in credit card fraud detection. Intelligent Systems and their Applications (IEEE), 14(3):67–74, 1999.
 37. Marc Rysman. An empirical analysis of payment card usage. The Journal of Industrial Economics, 55(1):1–36, 2007.
 38. Michael Szell, Sébastien Grauwin, and Carlo Ratti. Contraction of online response to major events. PloS ONE, 9(2):e89052, 2014.
 39. Morgan R Frank, Lewis Mitchell, Peter S Dodds, and Christopher M Danforth. Happiness and the patterns of life: A study of geolocated tweets. Scientific Reports, page 2625, 2013.
 40. Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Katakopoulos, and Carlo Ratti. Geo-located twitter as proxy for global mobility pattern. Cartography and Geographic Information Science, 41(3):260–271, 2014.
 41. Silvia Paldino, Iva Bojic, Stanislav Sobolevsky, Carlo Ratti, and Marta C González. Urban magnetism through the lens of geo-tagged photography. EPJ Data Science, 4(1):1–17, 2015.
 42. Maxime Lenormand, Thomas Louail, Oliva G. Cantu-Ros, Miguel Picornell, Ricardo Herranz, Juan Murillo Arias, Marc Barthelemy, Maxi San Miguel, and Jose J. Ramasco. Influence of sociodemographic characteristics on human mobility. arXiv:1411.7895, 2014.
 43. Thomas Louail, Maxime Lenormand, Oliva Garcia Cantu Ros, Migueal Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J. Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. Scientific Reports, 4:5276, 2014.
 44. Carlo Ratti, Stanislav Sobolevsky, Francesco Calabrese, Clio Andris, Jonathan Reades, Mauro Martino, Rob Claxton, and Steven H Strogatz. Redrawing the map of Great Britain from a network of human interactions. PLoS One, 5(12):1–6, 2010.

-
45. Sébastien Grauwin, Stanislav Sobolevsky, Simon Moritz, István Gódor, and Carlo Ratti. Towards a comparative science of cities: using mobile traffic records in New York, London and Hong Kong. arXiv:1406.4400, 2014.
 46. Tao Pei, Stanislav Sobolevsky, Carlo Ratti, Shih-Lung Shaw, Ting Li, and Chenghu Zhou. A new insight into land use classification based on aggregated mobile phone data. International Journal of Geographical Information Science, 28(9):1988–2007, 2014.
 47. S. Sobolevsky, I. Sitko, R. Tachet des Combes, B. Hawelka, J. Murillo Arias, and C. Ratti. Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. the case of residents and foreign visitors in spain. In Big Data (BigData Congress), 2014 IEEE International Congress on, Jun 27-Jul 2, Anchorage, AK, pages 136–143, 2014.
 48. A Noulas, S Scellato, R Lambiotte, M Pontil, and Cecilia Mascolo. A tale of many cities: Universal patterns in human urban mobility. Plos ONE, 7(5):e37027, 2012.
 49. Kevin Kung, Kael Greco, Stanislav Sobolevsky, and Carlo Ratti. Exploring universal patterns in human home/work commuting from mobile phone data. PLoS ONE, 9(6):e96180, 2014.
 50. Stanislav Sobolevsky, Emanuele Massaro, Iva Bojic, Juan Murillo Arias, and Carlo Ratti. Predicting regional economic indices using big data of individual bank card transactions. arXiv preprint arXiv:1506.00036, 2015.
 51. J. Lane, V. Stodden, S. Bender, and H. Nissenbaum. Privacy, big data, and the public good. Cambridge University Press, 2014.
 52. D. Christin, A. Reinhardt, S. S. Kanhere, and M. Hollick. A survey on privacy in mobile participatory sensing applications. Journal of Systems and Software, 84:1928–1946, 2011.
 53. F. Bélanger and R. E. Crossler. Privacy in the digital age: a review of information privacy research in information systems. Mis Quarterly, 35:1017–1042, 2011.
 54. I. Krontiris, F. C. Freiling, and T. Dimitriou. Location privacy in urban sensing networks: research challenges and directions [security and privacy in emerging wireless networks. Wireless Communications, IEEE, 17:30–35, 2010.
 55. Richard M. Walker and Rhys Andrews. Local government management and performance: A review of evidence. Journal of Public Administration Research and Theory, 2013.
 56. Constantine Kontokosta. The price of victory. the impact of the olympic games on residential real estate markets. Urban Studies, 49(5):961–978, 2012.
 57. Chris Hans. Bayesian lasso regression. Biometrika, 96(4):835–845, 2009.
 58. Simon Haykin. Neural networks and learning machines. 2009.
 59. Federico Girosi, Michael Jones, and Tomaso Poggio. Regularization theory and neural networks architectures. Neural computation, 7:219–269, 1995.
 60. Yaochu Jin, Tatsuya Okabe, and Bernhard Sendhoff. Neural network regularization and ensembling using multi-objective evolutionary algorithms. Evolutionary Computation, Congress on. Vol. 1. IEEE, 1, 2004.

-
61. Andy Liaw and Matthew Wiener. Classification and regression by randomforest. R news, 2(3):18–22, 2002.
 62. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.
 63. P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. Machine Learning, 63(1):3–42, 2006.

S1 Text. Unsupervised model and cluster number selection

Let S be a set of observations (locations in our data set), and a clustering U on S is a way of partitioning S into non-overlapping subsets U_1, U_2, \dots, U_k . We will investigate how well the model performs with different number of clusters, i.e. different k .

Here we choose K-means clustering with four clusters as our basic model. We made this decision based on the two clustering evaluation methods: Silhouette method and Elbow method^[1].

S1.1 Silhouette method

Silhouette is a commonly used method of interpretation and validation of consistency within clusters of data. It was first described by Peter J. Rousseeuw in 1986^[3] and it measures how similar an object is to its own cluster (internal relation) compared to other clusters (external relation). The Silhouette score ranges from -1 to 1, where higher value indicates better match to its own cluster and, at the same time, poorer matched to neighboring clusters—hence, higher Silhouette score means a better model overall as it highlights the distinctions among clusters.

The Silhouette value can be calculated with any distance metric, such as the Euclidean distance we applied here. We have run the tests for all three cities. The results are plotted on Figure S1.1. We can see that:

- New York City: Models with 2, 3, and 4 clusters seem closely comparable and outperform the rest;
- Boston: Models with 2 and 4 clusters have the highest Silhouette scores;
- Chicago: Silhouette score appears to be decreasing as the number of clusters rises, the optimal choice is 2.

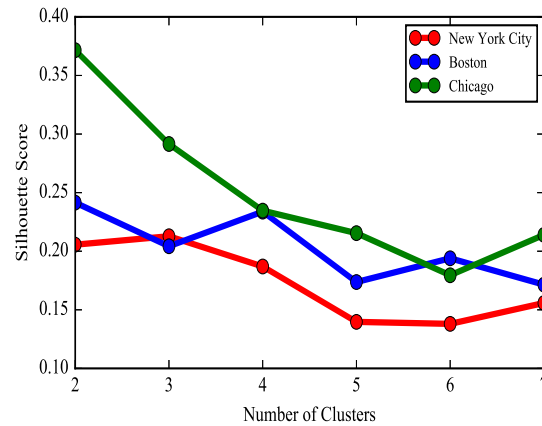


Figure S1.1. Silhouette method

This observation tells us two things:

- Models with 2, 3, and 4 clusters are generally better than others;
- 2-cluster model seems to be the best choice in terms of Silhouette's quantitative criteria.

Next we try Elbow method, another validation approach described in next subsection, before making final decisions.

S1.2 Elbow method

The Elbow method measures how "cost-efficient" a model is by looking at the percentage of variance explained as a function of the number of clusters. It searches for a balance between "more information" and "less complicated model". Intuitively, if we start from 1-cluster model (which is no processing at all, just leave them as a whole), adding another cluster should give more information about the data distinction, but one should stop when the marginal gain is insignificant compared to the cost. Then the number of clusters is chosen at this point^[5].

Equivalently, we can check the average sum of squared errors. Of course, we want our error as small as possible, and the error tends to decrease toward 0 as we increase the cluster number, k (the error is 0 when k is equal to the number of data points in the dataset, because then each data point is its own cluster, and there is no error between this point and the center of its cluster—that point itself). The goal is the same: search for the point where the marginal drop is no longer attractive beyond it.

The results are summarized in Figure S1.2:

- New York City: Obviously the error drops rapidly before 4 and then slows down after 4, so 4-cluster model is the best choice here;
- Chicago: Very similar to NYC, although the change is a bit mild and both - 3 and 4 - seem to be good choices;
- Boston: The trend does not provide any intuitive number of clusters to focus on.

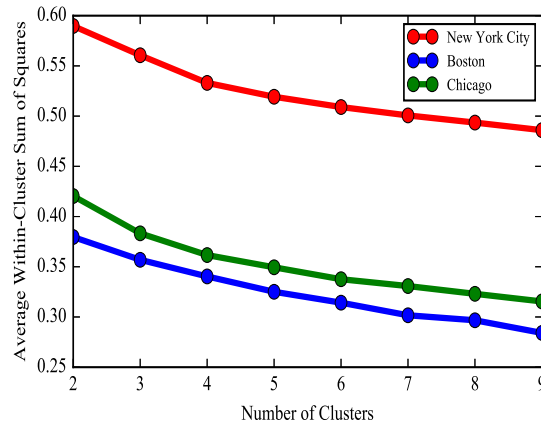


Figure S1.2. Elbow method

S1.3 Conclusion

To sum up, we have the following observations among three major cities in Table S1:

Since 4 is the only number that has appeared in all three rows, and clearly 4 clusters can reveal more details about the city structures than 2 or 3, we think that 4-cluster model may be the best overall choice. Choosing 4 instead of, say, 2, in our opinion, is a reasonable trade-off between having more clusters and still decent clustering quality.

	Silhouette	Elbow
NYC	3	4
Chicago	2	3, 4
Boston	2, 4	2

Table S1. Optimal choices based on each evaluation method

References

1. MacQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1967;1: Statistics: 281–297.
2. Vinh, N Xuan and Epps, Julien. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. Journal of Machine Learning Research 2010; 11:2837-2854
3. Rousseeuw P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20: 53–65.
4. Kaufman L, Rousseeuw P. Clustering by means of medoids. In: Dodge Y, editor. Statistical Data Analysis Based on the L1 Norm and Related Methods. North-Holland; 1987. p. 405–416.
5. Ketchen, J. David and Shook, L. Christopher. The application of cluster analysis in Strategic Management Research: An analysis and critique. Strategic Management Journal 1996; 17 (6): 441–458.
doi:10.1002/(SICI)1097-0266(199606)17:6;441::AID-SMJ819;3.0.CO;2-G.

S2 Text. Classification result based on 311 service requests timeline data

The 311 service request data is pretty rich and although types of requests considered in the paper provide an important and useful perspective for spatial clustering and modeling socio-economic quantities, there are other interesting dimensions in the data to consider. In this supplementary paragraph, we provide an alternative approach to conduct the clustering analysis based on 311 service requests data. Instead of using types of 311 service requests, we consider their timeline, building our new data set by accumulating all types of 311 services requests during each hour of the week for each zip code area. Thus this new data set includes 168 features (activity distribution per hours within an average), for all the zip code areas within New York City.

Based on the new 168 dimensional feature space, we divide the zip code areas in NYC into four clusters using K-Means clustering algorithm, highlighting different temporal patterns in 311 service request activity. The clustering result is shown on the Figure S2.1, while the corresponding 311 service request timelines for different clusters — on the Figure S2.2.

Figure S2.2 shows that the timelines of 311 service requests among Clusters 1,2,3 are rather similar. But Cluster 4 can be distinguished since the service requests have a considerable spike after 8 PM each day and especially over the weekends, indicating evening-time activity in those areas, which largely include locations across Manhattan, which makes common sense. Further understanding this pattern might require

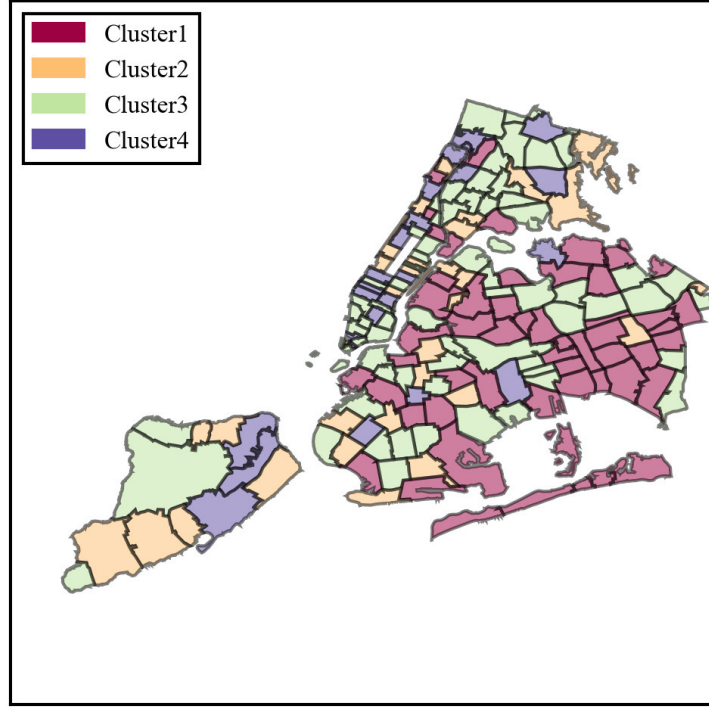


Figure S2.1. Classification of urban locations based on the timeline data of 311 request services

additional analysis of the service requests' types and other contextual information. However the overall difference between socio-economic factors among different clusters is much less significant than the result based on 311 service request type data shown in Figure 4. Therefore, for the purpose of the socio-economic analysis of this study we decided to stick to the service request types other the timeline.

S3 Test. Model selection, Cross-validation and Out of Sample R-squared

The results of the model selection are shown in Table 2. The following procedures are applied for each city.

Assuming we have data set X and labels y , where X is $n \times m$ matrix and y is $n \times 1$ vector. All the entries are real values.

Firstly, we collect all the possible models for training. The types of the models we consider include: Lasso, Neural Networks with regularization, Random Forest Regression, and Extra Trees Regression. We treat models with different hyper-parameters set as different models even they are belong to the same model type.

Secondly, we randomly split the whole data set into training data and test data. We train each model's parameters on training set, and get the out of sample R-squared

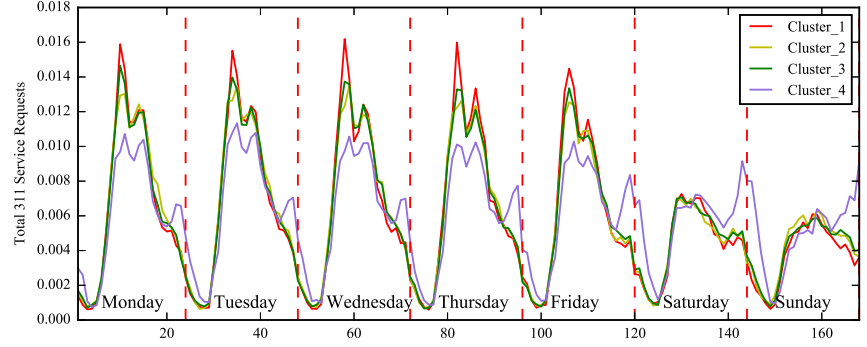


Figure S2.2. Hourly distribution of 311 service requests among clusters.

from the prediction result on testing set. We repeat this process ten times, and record the average R-squared for each model.

Finally, we report the largest out of sample R-squared among all the records(models) from second part and write it in Table 2.

References

1. Tibshirani, Robert. Regression Shrinkage and Selection via the lasso Journal of the Royal Statistical Society. Series B (methodological) 1996. 58 (1). Wiley: 267–88.
2. Girosi, Federico; Michael Jones; Tomaso Poggio. Regularization Theory and Neural Networks Architectures Neural Computation 1995. 7 (2): 219–269.
3. L.Breiman Random Forests Machine Learning, 45(1), 5-32, 2001.

S4 Test. Choice of scale among zip code, census tract and census block.

Consider the spatial granularity for the spatial aggregation of the New York City’s 311 service requests, choosing among the three options: zip code areas, census tract areas, and census block areas of New York City.

The primary goal for this scale selection is to find the right balance between the number of spatial units which will serve as observations for our model and the sparsity of the data. In Figure S4, we use x-axis for the number of total requests in each area (zip code level, census block level, etc), and for each given x show the number y of areas with request activity higher than x . We hope to find an appropriate scale such that it provides both adequate number of areas to analyze and abundant request activities to analyze per each area.

- Let’s start with Zip Code scale—only 178 total observations at hand, it’s too few to apply various machine learning algorithms for our research, despite most zip code areas have more than 500 activities in total.
- Census Block scale, on the other hand, offers more than 6000 areas in total. But for most of these areas (93%), total activities are less than 500.

- In comparison, Census Tract data set has 1367 observations with more than 500 requests, which seems to be a good balance addressing the issue of data sparsity as well as providing enough areas to analyze.

Hence we have selected Census Tract as the basic spatial scale for our research.

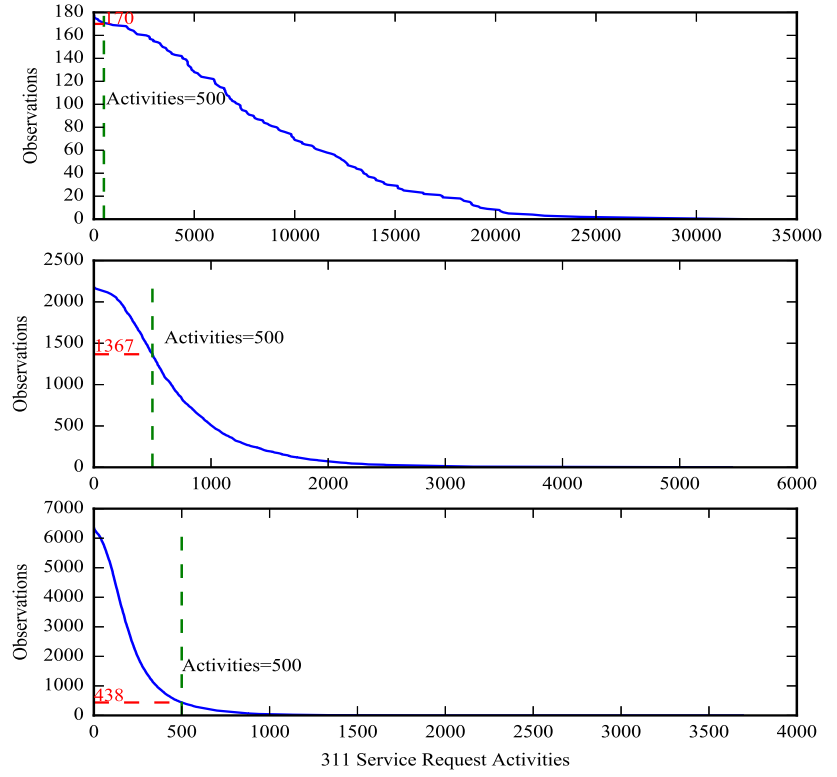


Figure S4. Number of areas vs Request Activity per Area